Abstract

Extractive reading comprehension systems can often locate the correct answer to a question in a context document, but they also tend to make unreliable guesses on questions for which the correct answer is not stated in the context. Existing datasets either focus exclusively on answerable questions, or use automatically generated unanswerable questions that are easy to identify. To address these weaknesses, we present SQuAD 2.0, the latest version of the Stanford Question Answering Dataset (SQuAD). SQuAD 2.0 combines existing SQuAD data with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD 2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD 2.0 is a challenging natural language understanding task for existing models: a strong neural system that gets 86% F1 on SQuAD 1.1 achieves only 66% F1 on SQuAD 2.0.

1 Introduction

Machine reading comprehension has become a central task in natural language understanding, fueled by the creation of many large-scale datasets (Hermann et al., 2015; Hewlett et al., 2016; Rajpurkar et al., 2016; Nguyen et al., 2016; Trischler et al., 2017; Joshi et al., 2017). In turn, these datasets have spurred a diverse array of model architecture improvements (Seo et al., 2016; Hu et al., 2017; Wang et al., 2017; Clark and Gardner, 2017; Huang et al., 2018). Recent work has even produced systems that surpass human-level exact match accuracy on the Stanford Question Answering Dataset (SQuAD), one of the most widely-used reading comprehension benchmarks (Rajpurkar et al., 2016).

Nonetheless, these systems are still far from true language understanding. Recent analysis shows that models can do well at SQuAD by learning context and type-matching heuristics (Weiszenborn et al., 2017), and that success on SQuAD does not ensure robustness to distracting sentences (Jia and Liang, 2017). One root cause of these problems is SQuAD’s focus on questions for which a correct answer is guaranteed to exist in the context document. Therefore, models only need to select the span that seems most related to the question, instead of checking that the answer is actually entailed by the text.

In this work, we construct SQuAD 2.0, a new dataset that combines answerable questions from the previous version of SQuAD (SQuAD 1.1)
with 53,775 new, unanswerable questions about the same paragraphs. Crowdworkers crafted these questions so that (1) they are relevant to the paragraph, and (2) the paragraph contains a plausible answer—something of the same type as what the question asks for. Two such examples are shown in Figure 1.

We confirm that SQuAD 2.0 is both challenging and high-quality. A state-of-the-art model achieves only 66.3% F1 score when trained and tested on SQuAD 2.0, whereas human accuracy is 89.5% F1, a full 23.2 points higher. The same model architecture trained on SQuAD 1.1 gets 85.8% F1, only 5.4 points worse than humans.

We also show that our unanswerable questions are more challenging than ones created automatically, either via distant supervision (Clark and Gardner, 2017) or a rule-based method (Jia and Liang, 2017). We release SQuAD 2.0 to the public as new version of SQuAD, and make it the primary benchmark on the official SQuAD leaderboard.2 We are optimistic that this new dataset will encourage the development of reading comprehension systems that know what they don’t know.

2 Desiderata

We first outline our goals for SQuAD 2.0. Besides the generic goals of large size, diversity, and low noise, we posit two desiderata specific to unanswerable questions:

Relevance. The unanswerable questions should appear relevant to the topic of the context paragraph. Otherwise, simple heuristics (e.g., based on word overlap) could distinguish answerable and unanswerable questions (Yih et al., 2013).

Existence of plausible answers. There should be some span in the context whose type matches the type of answer the question asks for. For example, if the question asks, “What company was founded in 1992?” then some company should appear in the context. Otherwise, type-matching heuristics could distinguish answerable and unanswerable questions (Weissenborn et al., 2017).

3 Existing datasets

Next, we survey existing reading comprehension datasets with these criteria in mind. We use the term “negative example” to refer to a context passage paired with an unanswerable question.

3.1 Extractive datasets

In extractive reading comprehension datasets, a system must extract the correct answer to a question from a context document or paragraph. The Zero-shot Relation Extraction dataset (Levy et al., 2017) contains negative examples generated with distant supervision. Levy et al. (2017) found that 65% of these negative examples do not have a plausible answer, making them easy to identify.

Other distant supervision strategies can also create negative examples. TriviaQA (Joshi et al., 2017) retrieves context documents from the web or Wikipedia for each question. Some documents do not contain the correct answer, yielding negative examples; however, these are excluded from the final dataset. Clark and Gardner (2017) generate negative examples for SQuAD by pairing existing questions with other paragraphs from the same article based on TF-IDF overlap; we refer to these as TFIDF examples. In general, distant supervision does not ensure the existence of a plausible answer in the retrieved context, and might also add noise, as the context might contain a paraphrase of the correct answer. Moreover, when retrieving from a small set of possible contexts, as in Clark and Gardner (2017), we find that the retrieved paragraphs are often not very relevant to the question, making these negative examples easy to identify.

The NewsQA data collection process also yields unanswerable questions, because crowdworkers write questions given only a summary of an article, not the full text (Trischler et al., 2017). Only 9.5% of their questions are unanswerable, making this strategy hard to scale. Of this fraction, we found that some are misannotated as unanswerable, and others are out-of-scope (e.g., summarization questions). Trischler et al. (2017) also exclude negative examples from their final dataset.

Jia and Liang (2017) propose a rule-based procedure for editing SQuAD questions to make them unanswerable. Their questions are not very diverse: they only replace entities and numbers with similar words, and replace nouns and adjectives with WordNet antonyms. We refer to these unanswerable questions as RULEBASED questions.

3.2 Answer sentence selection datasets

Sentence selection datasets test whether a system can rank sentences that answer a question higher
than sentences that do not. Wang et al. (2007) constructed the QASent dataset from questions in the TREC 8-13 QA tracks. Yih et al. (2013) showed that lexical baselines are highly competitive on this dataset. WikiQA (Yang et al., 2015) pairs questions from Bing query logs with sentences from Wikipedia. Like TF-IDF examples, these sentences are not guaranteed to have plausible answers or high relevance to the question. The dataset is also limited in scale (3,047 questions, 1,473 answers).

### 3.3 Multiple choice datasets
Finally, some datasets, like MCTest (Richardson et al., 2013) and RACE (Lai et al., 2017), pose multiple choice questions, which can have a “none of the above” option. In practice, multiple choice options are often unavailable, making these datasets less suited for training user-facing systems. Multiple choice questions also tend to be quite different from extractive ones, with more emphasis on fill-in-the-blank, interpretation, and summarization (Lai et al., 2017).

### 4 SQuAD 2.0
We now describe our new dataset, which we constructed to satisfy both the relevance and plausible answer desiderata from Section 2.

4.1 Dataset creation
We employed crowdworkers on the Daemo crowdsourcing platform (Gaikwad et al., 2015) to write unanswerable questions. Each task consisted of an entire article from SQuAD 1.1. For each paragraph in the article, workers were asked to pose up to five questions that were impossible to answer based on the paragraph alone, while referencing entities in the paragraph and ensuring that a plausible answer is present. As inspiration, we also showed questions from SQuAD 1.1 for each paragraph; this further encouraged unanswerable questions to look similar to answerable ones. Workers were asked to spend 7 minutes per paragraph, and were paid $10.50 per hour. Screenshots of our interface are shown in Appendix A.1.

We removed questions from workers who wrote 25 or fewer questions on that article; this filter helped remove noise from workers who had trouble understanding the task, and therefore quit before completing the whole article. We applied this filter to both our new data and the existing answerable questions from SQuAD 1.1. To generate train, development, and test splits, we used the same partition of articles as SQuAD 1.1, and combined the existing data with our new data for each split. For the SQuAD 2.0 development and test sets, we removed articles for which we did not
Table 2: Dataset statistics of SQuAD 2.0, compared to the previous SQuAD 1.1.

<table>
<thead>
<tr>
<th></th>
<th>SQuAD 1.1</th>
<th>SQuAD 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Train</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total examples</td>
<td>87,599</td>
<td>130,319</td>
</tr>
<tr>
<td>Negative examples</td>
<td>0</td>
<td>43,498</td>
</tr>
<tr>
<td>Total articles</td>
<td>442</td>
<td>442</td>
</tr>
<tr>
<td>Articles with negatives</td>
<td>0</td>
<td>285</td>
</tr>
<tr>
<td><strong>Development</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total examples</td>
<td>10,570</td>
<td>11,873</td>
</tr>
<tr>
<td>Negative examples</td>
<td>0</td>
<td>5,945</td>
</tr>
<tr>
<td>Total articles</td>
<td>48</td>
<td>35</td>
</tr>
<tr>
<td>Articles with negatives</td>
<td>0</td>
<td>35</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total examples</td>
<td>9,533</td>
<td>8,862</td>
</tr>
<tr>
<td>Negative examples</td>
<td>0</td>
<td>4,332</td>
</tr>
<tr>
<td>Total articles</td>
<td>46</td>
<td>28</td>
</tr>
<tr>
<td>Articles with negatives</td>
<td>0</td>
<td>28</td>
</tr>
</tbody>
</table>

4.2 Human accuracy

To confirm that our dataset is clean, we hired additional crowdworkers to answer all questions in the SQuAD 2.0 development and test sets. In each task, we showed workers an entire article from the dataset. For each paragraph, we showed all associated questions; unanswerable and answerable questions were shuffled together. For each question, workers were told to either highlight the answer in the paragraph, or mark it as unanswerable. Workers were told to expect every paragraph to have some answerable and some unanswerable questions. They were asked to spend one minute per question, and were paid $10.50 per hour.

To reduce crowdworker noise, we collected multiple human answers for each question and selected the final answer by majority vote, breaking ties in favor of answering questions and preferring shorter answers to longer ones. On average, we collected 4.8 answers per question. We note that for SQuAD 1.1, Rajpurkar et al. (2016) evaluated a single human’s performance; therefore, they likely underestimate human accuracy.

5 Experiments

5.1 Models

We evaluated three existing model architectures: the BiDAF-No-Answer (BNA) model proposed by Levy et al. (2017), and two versions of the DocumentQA No-Answer (DocQA) model from Clark and Gardner (2017), namely versions with and without ELMo (Peters et al., 2018). These models all learn to predict the probability that a question is unanswerable, in addition to a distribution over answer choices. At test time, models abstain whenever their predicted probability that a question is unanswerable exceeds some threshold. We tune this threshold separately for each model on the development set. When evaluating on the test set, we use the threshold that maximizes F1 score on the development set. We find this strategy does slightly better than simply taking the argmax prediction, possibly due to the different proportions of negative examples at training and test time.

5.2 Main results

First, we trained and tested all three models on SQuAD 2.0, as shown in Table 3. Following Rajpurkar et al. (2016), we report average exact match and F1 scores. The best model, DocQA + ELMo, achieves only 66.3 F1 on the test set, 23.2 points lower than the human accuracy of 89.5 F1. Note that a baseline that always abstains gets 48.9 test F1; existing models are closer to this baseline than they are to human performance. Therefore, we see significant room for model improvement on this task. We also compare with reported test numbers for analogous model architectures on SQuAD 1.1. There is a much larger gap between humans and machines on SQuAD 2.0 compared to SQuAD 1.1, which confirms that SQuAD 2.0 is a much harder dataset for existing models.

3 For negative examples, abstaining receives a score of 1, and any other response gets 0, for both exact match and F1.
<table>
<thead>
<tr>
<th>System</th>
<th>SQuAD 1.1 test EM</th>
<th>SQuAD 1.1 test F1</th>
<th>SQuAD 2.0 dev EM</th>
<th>SQuAD 2.0 dev F1</th>
<th>SQuAD 2.0 test EM</th>
<th>SQuAD 2.0 test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNA</td>
<td>68.0</td>
<td>77.3</td>
<td>59.8</td>
<td>62.6</td>
<td>59.2</td>
<td>62.1</td>
</tr>
<tr>
<td>DocQA</td>
<td>72.1</td>
<td>81.0</td>
<td>61.9</td>
<td>64.8</td>
<td>59.3</td>
<td>62.3</td>
</tr>
<tr>
<td>DocQA + ELMo</td>
<td>78.6</td>
<td>85.8</td>
<td>65.1</td>
<td>67.6</td>
<td>63.4</td>
<td>66.3</td>
</tr>
<tr>
<td>Human</td>
<td>82.3</td>
<td>91.2</td>
<td>86.3</td>
<td>89.0</td>
<td>86.9</td>
<td>89.5</td>
</tr>
<tr>
<td>Human–Machine Gap</td>
<td>3.7</td>
<td>5.4</td>
<td>21.2</td>
<td>21.4</td>
<td>23.5</td>
<td>23.2</td>
</tr>
</tbody>
</table>

Table 3: Exact Match (EM) and F1 scores on SQuAD 1.1 and 2.0. The gap between humans and the best tested model is much larger on SQuAD 2.0, suggesting there is a great deal of room for model improvement.

<table>
<thead>
<tr>
<th>System</th>
<th>SQuAD 1.1 + TF IDF EM</th>
<th>SQuAD 1.1 + TF IDF F1</th>
<th>SQuAD 1.1 + RULEBASED EM</th>
<th>SQuAD 1.1 + RULEBASED F1</th>
<th>SQuAD 2.0 dev EM</th>
<th>SQuAD 2.0 dev F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNA</td>
<td>72.7</td>
<td>76.6</td>
<td>80.1</td>
<td>84.8</td>
<td>59.3</td>
<td>62.0</td>
</tr>
<tr>
<td>DocQA</td>
<td>75.6</td>
<td>79.2</td>
<td>80.8</td>
<td>84.8</td>
<td>61.9</td>
<td>64.8</td>
</tr>
<tr>
<td>DocQA + ELMo</td>
<td>79.4</td>
<td>83.0</td>
<td>85.7</td>
<td>89.6</td>
<td>65.1</td>
<td>67.6</td>
</tr>
</tbody>
</table>

Table 4: Exact Match (EM) and F1 scores on the SQuAD 2.0 development set, compared with SQuAD 1.1 with two types of automatically generated negative examples. SQuAD 2.0 is more challenging for current models.

5.3 Automatically generated negatives

Next, we investigated whether automatic ways of generating negative examples can also yield a challenging dataset. We trained and tested all three model architectures on SQuAD 1.1 augmented with either TF IDF or RULEBASED examples. To ensure a fair comparison with SQuAD 2.0, we generated training data by applying TF IDF or RULEBASED only to the 285 articles for which SQuAD 2.0 has unanswerable questions. We tested on the articles and answerable questions in the SQuAD 2.0 development set, adding unanswerable questions in a roughly one-to-one ratio with answerable ones. These results are shown in Table 4. The highest score on SQuAD 2.0 is 15.4 F1 points lower than the highest score on either of the other two datasets, suggesting that automatically generated negative examples are much easier for existing models to detect.

5.4 Plausible answers as distractors

Finally, we measured how often systems were fooled into answering the plausible but incorrect answers provided by crowdworkers for our unanswerable questions. For both computer systems and humans, roughly half of all wrong answers on unanswerable questions exactly matched the plausible answers. This suggests that the plausible answers do indeed serve as effective distractors. Full results are shown in Appendix A.2.

6 Discussion

SQuAD 2.0 forces models to understand whether a paragraph entails that a certain span is the answer to a question. Similarly, recognizing textual entailment (RTE) requires systems to decide whether a hypothesis is entailed by, contradicted by, or neutral with respect to a premise (Marelli et al., 2014; Bowman et al., 2015). Relation extraction systems must understand when a possible relationship between two entities is not entailed by the text (Zhang et al., 2017).

Jia and Liang (2017) created adversarial test examples that fool models trained on SQuAD 1.1. However, models that are trained on similar examples are not easily fooled by their method. In contrast, the adversarial examples in SQuAD 2.0 are difficult even for models trained on examples from the same distribution.

In conclusion, we have presented SQuAD 2.0, a challenging, diverse, and large-scale dataset that forces models to understand when a question cannot be answered given the context. We are optimistic that SQuAD 2.0 will encourage the development of new reading comprehension models that know what they don’t know, and therefore understand language at a deeper level.

Reproducibility. All code, data, experiments are available on the CodaLab platform at https://bit.ly/2rDHbGy.

Acknowledgments. We would like to thank the anonymous reviewers, Arun Chaganty, Peng Qi, and Sharon Zhou for their constructive feedback. We are grateful to Durim Morina and Michael Bernstein for their help with the Daemo platform. This work was supported by funding from Facebook. R.J. is supported by an NSF Graduate Research Fellowship under Grant No. DGE-114747.
References


A Supplementary material

A.1 Crowdsourcing details

Figure 2 shows the instructions that crowdworkers were given at the beginning of each question writing task. Figure 3 shows the interface they used to write unanswerable questions for each paragraph. In the interface, workers first write an unanswerable question, then highlight a plausible answer in the paragraph.

A.2 Plausible answers as distractors

As mentioned in Section 5.4, we measured how often systems were fooled into answering the plausible answers provided by crowdworkers for our unanswerable questions. For each system, we first isolated their false positive errors—cases where they predicted an answer to an unanswerable question—on the development set. Within this set of examples, we measured exact match and F1 scores between the system predictions and plausible answers. These numbers are shown in Table 5. Plausible answers account for roughly half of the false positive errors made by each of the computer systems, as well as by human answerers. We conclude that the plausible answers in our dataset do indeed serve their purpose of being distracting spans that could be mistaken for the correct answer.

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNA</td>
<td>48.6</td>
<td>63.0</td>
</tr>
<tr>
<td>DocQA</td>
<td>55.0</td>
<td>69.9</td>
</tr>
<tr>
<td>DocQA +ELMo</td>
<td>54.9</td>
<td>69.2</td>
</tr>
<tr>
<td>Human</td>
<td>46.4</td>
<td>60.6</td>
</tr>
</tbody>
</table>

Table 5: Exact match (EM) and F1 scores between system predictions and plausible answers, in cases where the system made a false positive error.

A.3 Question-only classification

To see if our unanswerable questions are easily distinguishable from answerable ones, we built a classifier to predict whether a question is answerable or unanswerable, looking only at the question itself. We used a linear SVM with unigram, bigram, and trigram features from the question, as well as an indicator for the number of words in the question. We trained on all questions in the SQuAD 2.0 training set, and got 61.0% binary accuracy on the development set, after a hyperparameter search on the regularization strength. We conclude that there is some signal in the questions alone that can distinguish answerable and unanswerable questions, but it is fairly weak.

We also inspected the trained model’s weights, and found that it learned that negation words (e.g., “never,” “n’t,” and “not”) and antonyms of words that often denote important events (e.g., “least,” “smallest,” and “last”, antonyms of “most”, “largest”, and “first”) were associated with unanswerable questions.
**Ask Impossible Reading Comprehension Questions**

**Instructions**

In this article about Geology, you will be asked to pose and answer reading comprehension questions based on the paragraph. There is a twist! The question you pose must be impossible to answer based on the paragraph alone, but should be about the same topic and same people/places/things! Additionally, the paragraph must contain a phrase/word that seems like a plausible answer to the question. Read each paragraph, pose an impossible question, and then highlight a phrase from the paragraph that looks like a plausible (but of course, incorrect) answer. We'll clarify the task with the help of an example below.

**Estimated Time For Task Completion - 3.2 hours**

This article consists of 25 paragraphs. We recommend a time of 7 minutes per paragraph. Submit each paragraph after you are done to save partial progress. Feel free to take breaks -- if you come back to the task, you do not need to resubmit paragraphs already submitted in an earlier session. After completing all paragraphs, click the submit task button at the end of the page.

---

The processes of decentralization redefine structures, procedures and practices of governance to be closer to the citizenry and to make them more aware of the costs and benefits; it is not merely a movement of power from the central to the local government. According to the United Nations Development Programme it is "more than a process, it is a way of life and a state of mind." The report provides a chart-formatted framework for defining the application of the concept 'decentralization' describing and elaborating on the "who, what, when, where, why and how" factors in any process of decentralization.

**Task Tutorial**

1. On the left, you’ll see a reading passage and ‘prompt questions’ underneath it. First read the passage, and skim over the questions.
2. Now, based on what you’ve read in the passage, your task is to come up with questions that don’t have a correct answer in the passage, but have feasible answers.
3. Start by picking a question from the prompt questions. For the purposes of this example, let’s pick “What is decentralization according to the United Nations Development Programme?”
4. Note that this question does have an answer in the passage. We’re going to modify it so that it doesn’t have an answer. For instance, we can modify the question to “What is decentralization according to the local government?” Note that this question doesn’t have an answer in the passage, but still contains entities present in the passage such as local government.
5. You will also be asked to pick a plausible answer for our question. This is an answer that looks possibly correct if someone hadn’t read the passage. You select the plausible answer by highlighting a segment of the passage. For our example question, we would highlight “it is more than a process, it is a way of life and a state of mind.”
6. Let’s come up with another example. This time, we will use the inspiration question “What is a way of life and state of mind, more than a

Figure 2: The instructions shown to crowdworkers at the beginning of each question writing task.
Figure 3: The interface crowdworkers used to write unanswerable questions and annotate plausible answers.